

10 July 2024

AI Risks and audit: Cybersecurity & privacy risks

Ioanna Topa,
IT Auditor,
DATA team



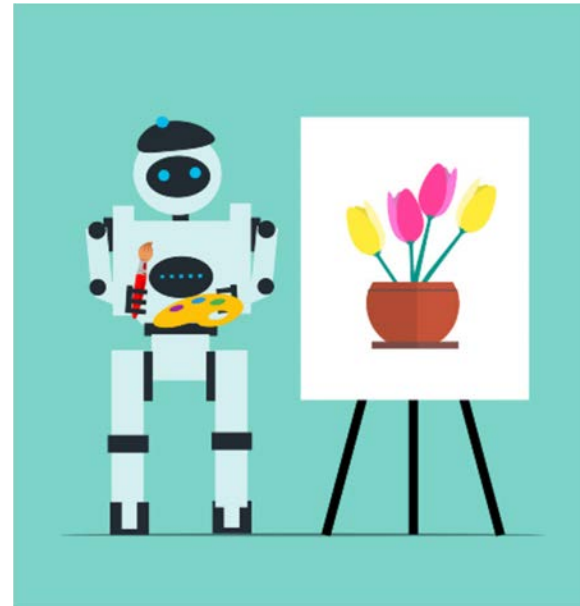
Outline

- What is an AI system?
- Information security
- AI - cybersecurity practices
- AI Risks for cybersecurity
- Everyday use of AI – cybersecurity and privacy risks
- AI lifecycle
- Characteristics of trustworthy systems
- Governance, Risk, Compliance
- ISO 42001
- EU AI Act



What is Artificial Intelligence?

- AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments

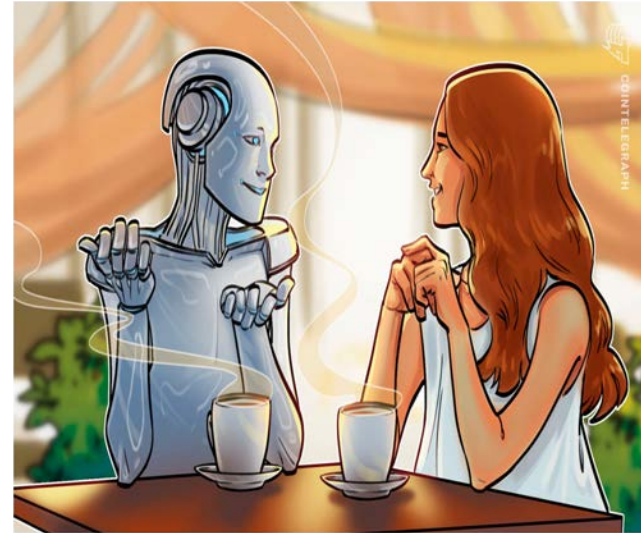


EU AI Act, <https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf>



Every day use of AI

- Unlocking the mobile phone with face ID
- Social networks and internet
 - Recommended songs
 - Recommended movies
 - Featured Ads
 - Recommended articles
- Emails – separating incoming messages into spam and junk mail
- Use of GPS
- Electrical devices
- Autonomous vehicles
- E-banking applications
- Health applications and systems



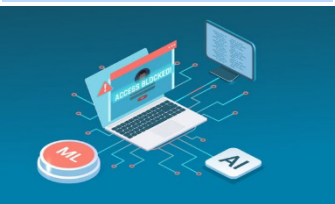
Information security

- Basic principles:
 - Confidentiality
 - Integrity
 - Availability



- Privacy issues:
 - Data protection





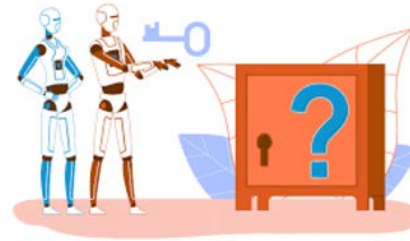
AI: Improved cybersecurity tools & practices

- Improved tools and techniques
 - For user authentication, instead of using complex passwords, e.g. with the use of biometrics such as face ID
 - For the identification of viruses
 - For the identification of phishing emails
 - For producing secure code(e.g. with ChatGPT)
 - For the identification of vulnerabilities
 - For the identification and resolution of security incidents (e.g. via advanced analysis of log files)
 - For performing penetration tests



Risks of AI in cybersecurity

- Passwords can easily be cracked
- Hackers can use ChatGPT to produce malicious code and malware, ransomware, exploits
- Launch hacking attacks
- Social engineering:
 - Sophisticated phishing emails
 - Deep fakes
 - Voice cloning
 - Generate synthetic identities
- Disinformation
- User profiling



Time it takes using AI to crack your password [2023]

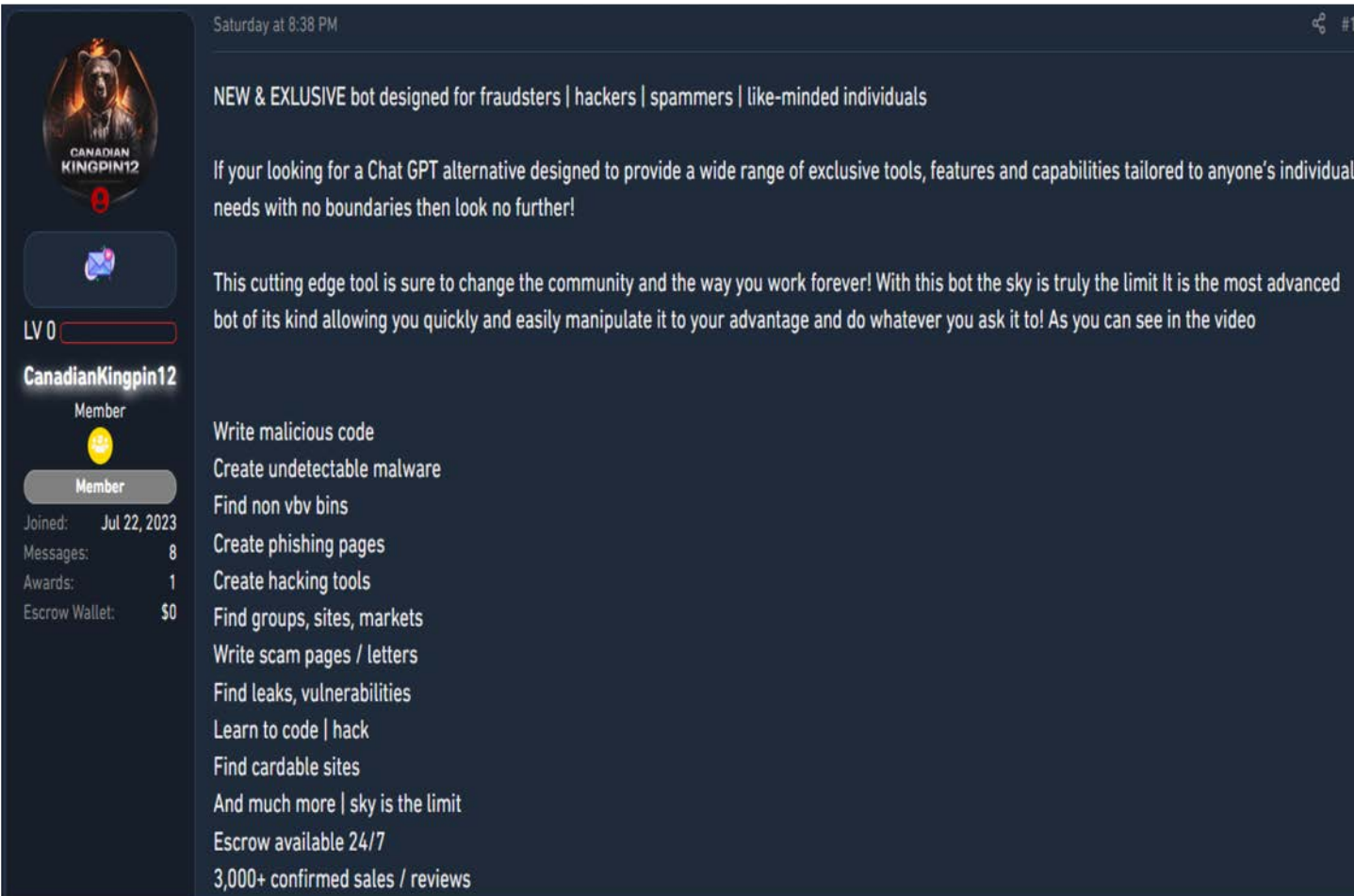
# OF CHARACTER	Numbers Only	Lowercase Letters	Lowercase Upper & Letters	Numbers, Upper & Lowercase Letters	Numbers, Upper & Lowercase Letters, Symbols
4	Instantly	Instantly	Instantly	Instantly	Instantly
5	Instantly	Instantly	Instantly	Instantly	Instantly
6	Instantly	Instantly	Instantly	Instantly	4 Seconds
7	Instantly	Instantly	22 Seconds	42 Seconds	6 Minutes
8	Instantly	3 Seconds	19 Minutes	48 Minutes	7 Hours
9	Instantly	1 Minutes	11 Hours	2 Days	2 Weeks
10	Instantly	1 Hours	4 Weeks	6 Months	5 Years
11	Instantly	23 Hours	4 Years	38 Years	356 Years
12	25 Seconds	3 Weeks	289 Years	2K Years	30K Years
13	3 Minutes	11 Months	16K Years	91K Years	2M Years
14	36 Minutes	49 Years	827K Years	9M Years	187M Years
15	5 Hours	890 Years	47M Years	613M Years	14Bn Years
16	2 Days	23K Years	2Bn Years	26Bn Years	1Tn Years
17	3 Weeks	812K Years	539.72M Years	2Tn Years	95Tn Years
18	10 Months	22M Years	7.23Bn Years	96Tn Years	6Qn Years

<https://www.zdnet.com/article/how-an-ai-tool-could-crack-your-passwords-in-seconds/>



Chat GPT for malware

In the wake of WormGPT, a ChatGPT clone trained on malware-focused data, a new generative artificial intelligence hacking tool called FraudGPT has emerged.



The screenshot shows a Discord chat window with a dark theme. On the left is a user profile for 'CanadianKingpin12', a member since July 22, 2023, with 8 messages and 1 award. The chat message, dated Saturday at 8:38 PM, advertises a bot designed for fraudsters, hackers, and spammers. The message text is as follows:

NEW & EXCLUSIVE bot designed for fraudsters | hackers | spammers | like-minded individuals

If your looking for a Chat GPT alternative designed to provide a wide range of exclusive tools, features and capabilities tailored to anyone's individual needs with no boundaries then look no further!

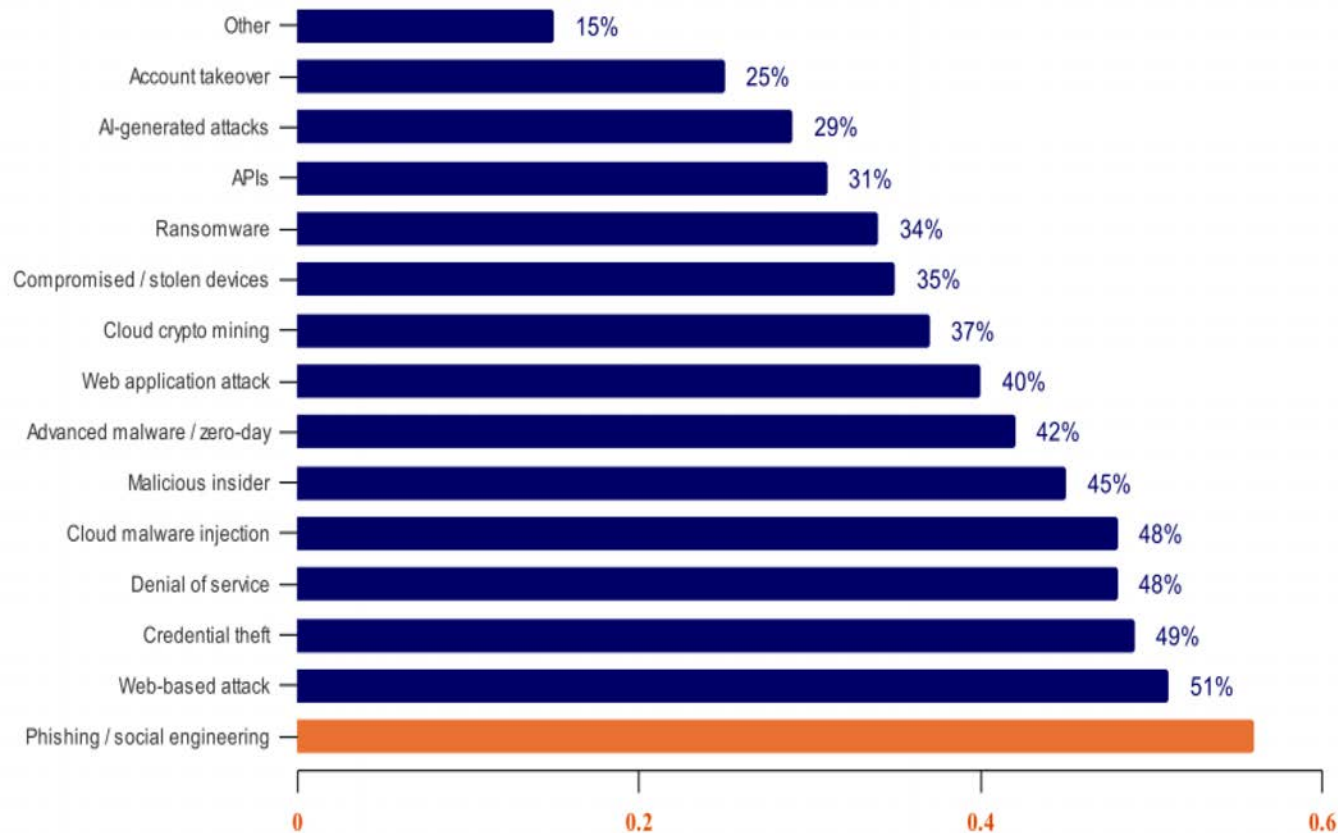
This cutting edge tool is sure to change the community and the way you work forever! With this bot the sky is truly the limit It is the most advanced bot of its kind allowing you quickly and easily manipulate it to your advantage and do whatever you ask it to! As you can see in the video

Write malicious code
Create undetectable malware
Find non vbv bins
Create phishing pages
Create hacking tools
Find groups, sites, markets
Write scam pages / letters
Find leaks, vulnerabilities
Learn to code | hack
Find cardable sites
And much more | sky is the limit
Escrow available 24/7
3,000+ confirmed sales / reviews

<https://www.bleepingcomputer.com/news/security/cybercriminals-train-ai-chatbots-for-phishing-malware-attacks/>



Cyber security attacks from AI-adopting organisations in the field of cybersecurity



Ponemon Institute surveyed 641 IT and security practitioners in organizations that are at some stage of AI adoption for cybersecurity

(State of AI in Cybersecurity Report 2024, Ponemon)



Example of phishing email created by ChatGPT (October 2023)

GPT-4

ML Hello, can you write me an example of a phishing email?

Sure, here's a common example of a phishing email. However, please note this is only for educational purposes and it's important not to use this to engage in fraudulent activities. It's crucial to maintain ethical standards when communicating online.



Subject: Your Account Has Been Compromised!

Dear Customer,

We noticed several unusual activities on your account. This includes multiple login attempts from different locations that we suspect to be fraudulent. Your account has been temporarily locked to ensure your safety.

To confirm your identity and unlock your account, kindly click the link below:

[Insert malicious link here]

Ensure you fill in the correct details to avoid permanent closure of your account.

If you did not request for this, please disregard this email. Your account will be safe.

Thanks for your understanding,

Customer Support Service

[Your Bank's Name]

Regenerate

Send a message



ML thank you, can you translate it into greek?

Of course, here is that same phishing email translated into Greek:



Θέμα: Ο Λογαριασμός σας Έχει Υποστεί Παραβίαση!

Αγαπητέ Πελάτη,

Παρατηρήσαμε αρκετές ασυνήθιστες δραστηριότητες στον λογαριασμό σας. Αυτό περιλαμβάνει πολλαπλές προσπάθειες σύνδεσης από διάφορες τοποθεσίες που υποψιαζόμαστε ότι είναι απατηλές. Ο λογαριασμός σας έχει προσωρινά κλειδωθεί για να διασφαλιστεί η ασφάλειά σας.

Για να επιβεβαιώσετε την ταυτότητά σας και να ξεκλειδώσετε τον λογαριασμό σας, παρακαλώ κάντε κλικ στον παρακάτω σύνδεσμο:

[Εισάγετε εδώ τον επιβλαβή σύνδεσμο]

Βεβαιωθείτε ότι συμπληρώνετε τα σωστά στοιχεία για να αποφύγετε τη μόνιμη κλείση του λογαριασμού σας.

Εάν δεν ζητήσατε αυτό, παρακαλούμε να αγνοήσετε αυτό το email. Ο λογαριασμός σας θα είναι ασφαλής.

Ευχαριστούμε για την κατανόησή σας,

Υπηρεσία Υποστήριξης Πελατών

[Το Όνομα της Τράπεζάς σας]




EUROPEAN
COURT
OF AUDITORS



Authentication


- Two-factor authentication

Something you
KNOW



Password or phrase
PIN

Something you
HAVE



Code from app or SMS
Push notification
USB token

Something you
ARE



Finger or thumb print
Face scan
Iris scan

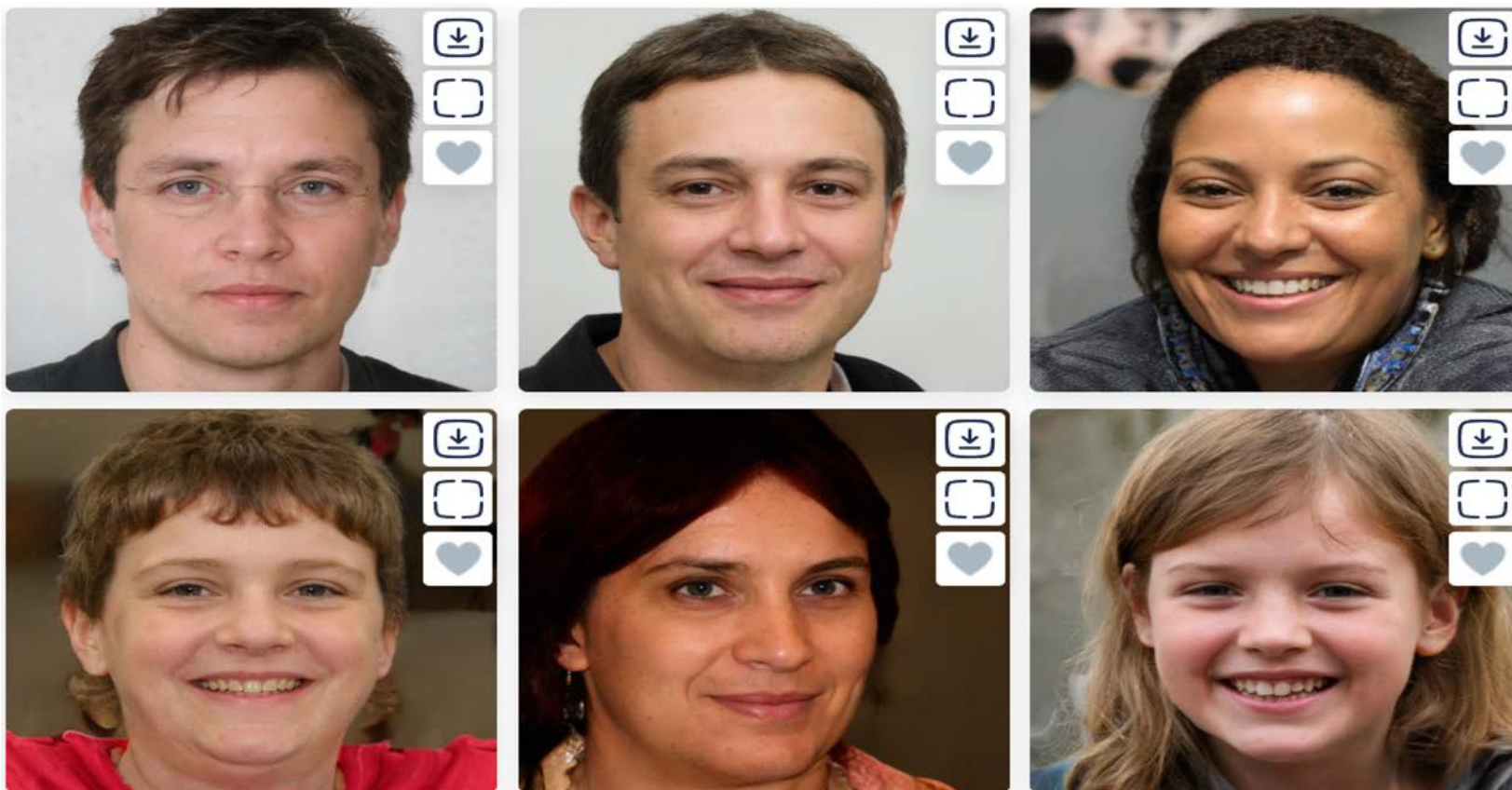


Biometrics – face recognition

- Face recognition with biometrics such as face ID
 - **Security risks**
 - Hackers use AI to bypass biometric security (<https://www.itweb.co.za/article/hackers-use-ai-to-bypass-biometrics-security/LPp6VMrBgVoMDKQz>)
 - Someone could manipulate the code so that the system cannot recognise faces correctly, e.g. a criminal could manipulate the code so that his face is not recognised by face recognition systems at the airport
 - **Privacy risks** – real-time face recognition



These people do not exist



https://thispersonnotexist.org/#google_vignette




EUROPEAN
COURT
OF AUDITORS



AI Real-Time Voice Cloning

 **HitPaw Voice Changer**

DOWNLOAD

 **HitPaw Voice Changer** *AI-Powered*

AI Real-Time Voice Cloning

- ✓ Sound as a variety of celebrities makes gaming or live-streaming more fun!
- ✓ Support uploading your own vocals into different celebrities and animated characters
- ✓ Various soundboards suitable for live-streaming, make it's experience so much fun.
- ✓ Make your virtual characters feel more alive and realistic.



Deepfakes

- Deepfake – A deepfake is the act of maliciously replacing real images and videos with fabricated ones to perform information manipulation.
- Deepfakes are usually a form of impersonation and can be achieved by changing one person's face with AI in order to look like another person and to convince others that he/she is the person they are trying to imitate.



- Security risks: Hackers could use deepfakes to get personal information from people
 - e.g. a deepfake hologram of a Chief Communications' officer of a cryptocurrency company during a Zoom videocall led the executives to reveal confidential information



Disinformation

- With the use of AI (e.g. generative AI) it is possible to create pictures, sounds, videos which are not real.
- These can lead to disinformation because they look real but they are fake.
- It is not easy to distinguish what is real and what is not.
- **Security risks:** There is a risk that a user reveals his username and password or credit card details to a hacker who had provided him with a fake picture or a fake audio message of a relative or of a close friend asking for help.



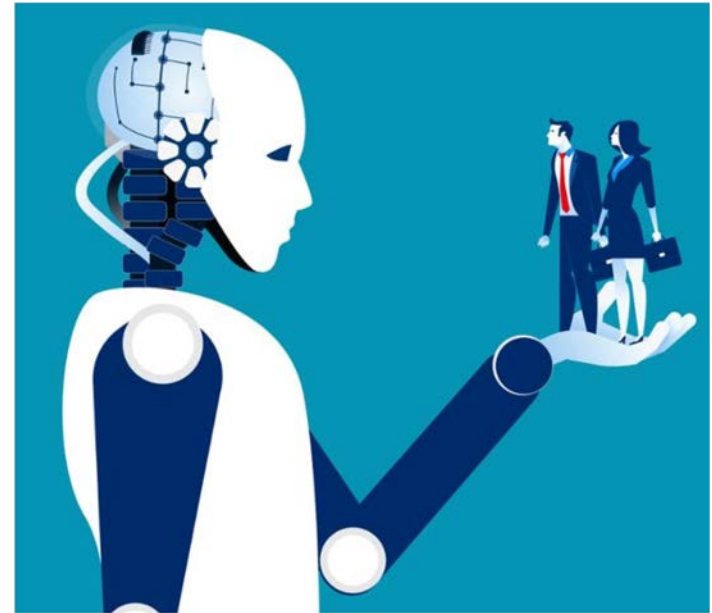
User profiling

- **User profiling**
 - The system stores information about the user, e.g. websites that this person has visited, history search. The algorithm provides the users with targeted advertisements, products and services to buy.
 - **Security risks**
 - Customised/targeted attacks
 - **Privacy risks**
 - Social scoring



Training data – Example of a tool for hiring

- There are systems using AI for selecting the cv of candidates based on key words.
- The data set used for training the algorithm plays an important role in providing assurance that the algorithm will give fair and unbiased results.
- **Ethical issues:**
 - In one case the algorithm was trained on data which included only men, because for that specific position, only men were previously hired.
 - As a result, the algorithm would select men for this position and not women, based on the data that were available. The selection was biased.



Training data – Example from AI algorithms for legal decisions

Legal issues

- The data set used for training the algorithm plays an important role in providing assurance that the algorithm will give up to date, fair and unbiased results.
- If the algorithm was trained on legal decisions from the past, then it is not possible to provide legal decisions that are up to date.
- One disadvantage for legal decisions generated with AI is that they lack in quality. It may be that the algorithm is trained on a large data set for a specific topic however, it is not possible to provide a customised legal decision.



Manipulation of AI Algorithms and training data among the top 10 Emerging Cybersecurity threats for 2030

10



Artificial Intelligence Abuse

Manipulation of AI algorithms and training data can be used to enhance nefarious activities such as the creation of disinformation and fake content, bias exploitation, collecting biometrics and other sensitive data, military robots and data poisoning.

- 1 Supply chain compromise of software dependencies**
More integrated components and services from third party suppliers and partners could lead to novel and unforeseen vulnerabilities with compromises on the supplier and customer side.
- 2 Advanced disinformation campaigns**
Deepfake attacks can manipulate communities for (geo)political reasons and for monetary gain.
- 3 Rise of digital surveillance authoritarianism/ loss of privacy**
Facial recognition, digital surveillance on internet platforms or digital identities data stores may become a target for criminal groups.
- 4 Human error and exploited legacy systems within cyber-physical ecosystems**
The fast adoption of IoT, the need to retrofit legacy systems and the ongoing skill shortage could lead to a lack of knowledge, training and understanding of the cyber-physical ecosystem, which can lead to security issues.
- 5 Targeted attacks enhanced by smart device data**
Through data obtained from internet connected smart devices, attackers can access information for tailored and more sophisticated attacks.
- 6 Lack of analysis and control of space-based infrastructure and objects**
Due to the intersections between private and public infrastructure in space, the security of these new infrastructures and technologies need to be investigated as a lack of understanding, analysis and control of space-based infrastructure can make it vulnerable to attacks and outages.
- 7 Rise of advanced hybrid threats**
Physical or offline attacks are evolving and becoming often combined with cyberattacks due to the increase of smart devices, cloud usage, online identities and social platforms.
- 8 Skill shortage**
Lack of capacities and competencies could see cybercriminal groups target organisations with the largest skills gap and the least maturity.
- 9 Cross border ICT service providers as a single point of failure**
ICT sector connecting critical services such as transport, electric grids and industry that provide services across borders are likely to be targeted by techniques such as backdoors, physical manipulation, and denial of service and weaponised during a future potential conflict.
- 10 Artificial Intelligence Abuse**
Manipulation of AI algorithms and training data can be used to enhance nefarious activities such as the creation of disinformation and fake content, bias exploitation, collecting biometrics and other sensitive data, military robots and data poisoning.



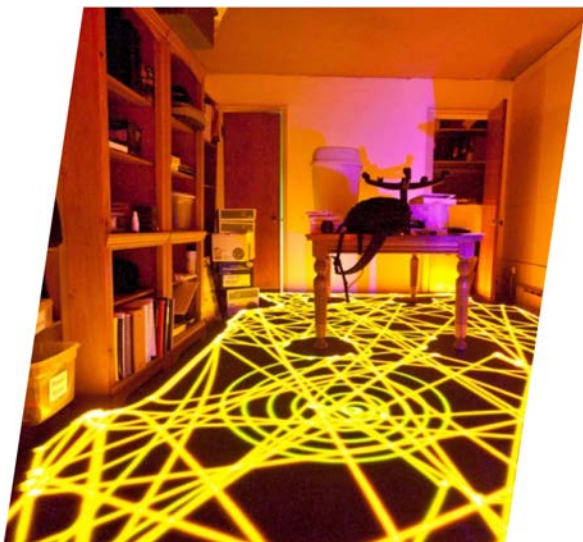
Everyday use of AI – Autonomous vehicle

- **Security risks**
 - What if someone hacks the car with a virus, e.g. there is a risk that he will be able to take control over and navigate the car.
- **Ethical issues**
 - If an obstacle appears in front of the car, by what criteria does the algorithm decide what to do?
 - If there are pedestrians on the sidewalk will it hit them?
 - Will the algorithm decide according to the age of the people?



Every day use of AI – Electric appliances

- Robot vacuum cleaner
- **Security risks**
 - A hacker exploits a vulnerability of the application and has access to the network of the house
- **Privacy risks**
 - The app has information about the architecture of the house
- Smart fridge
- **Security risks**
 - Someone could hack the refrigerator and order large quantities of products

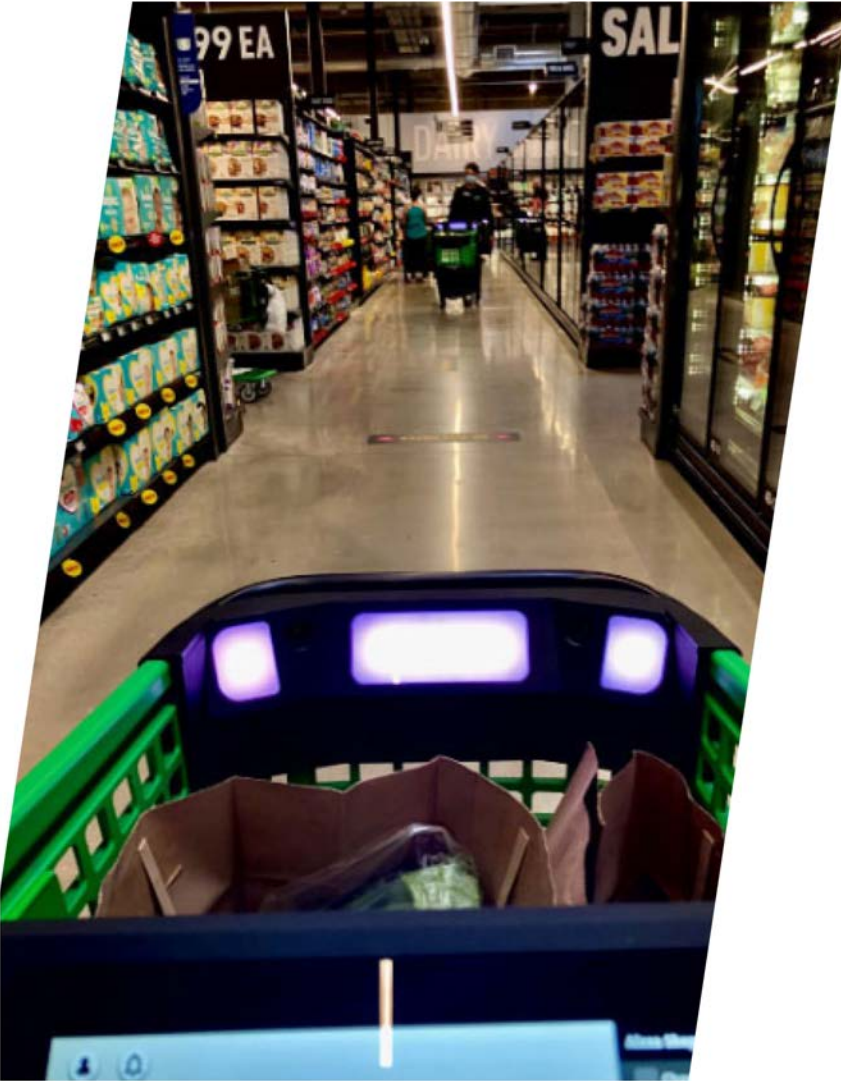


Every day use of AI - Supermarket(1/2)



<https://www.youtube.com/watch?v=NrmMk1Myrxc>



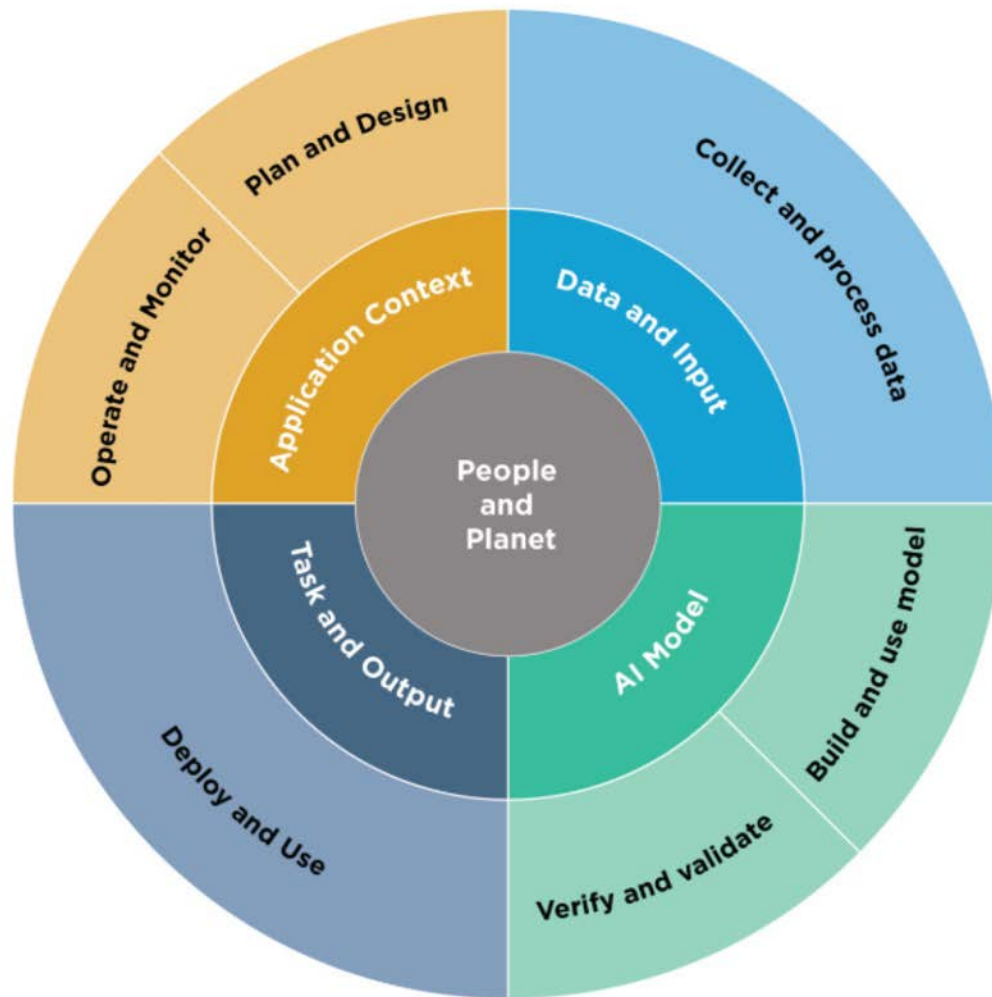


Every day use of AI – Supermarket (2/2)

- Supermarket where people can buy products without passing from the cashier. The payment is done automatically with the mobile phone.
- **Security risks**
 - If someone steals the phone, then he/she can make unauthorised purchases
- **Privacy risks**
 - Profiling, the application will know the diet habits of the consumers, what products he/she buys, and send targeted advertisements for products.
 - What will happen if the company decides to sell the application data?



Lifecycle and Key Dimensions of an AI System



Characteristics of trustworthy AI systems



Safe

Secure &
Resilient

Explainable &
Interpretable

Privacy
Enhanced

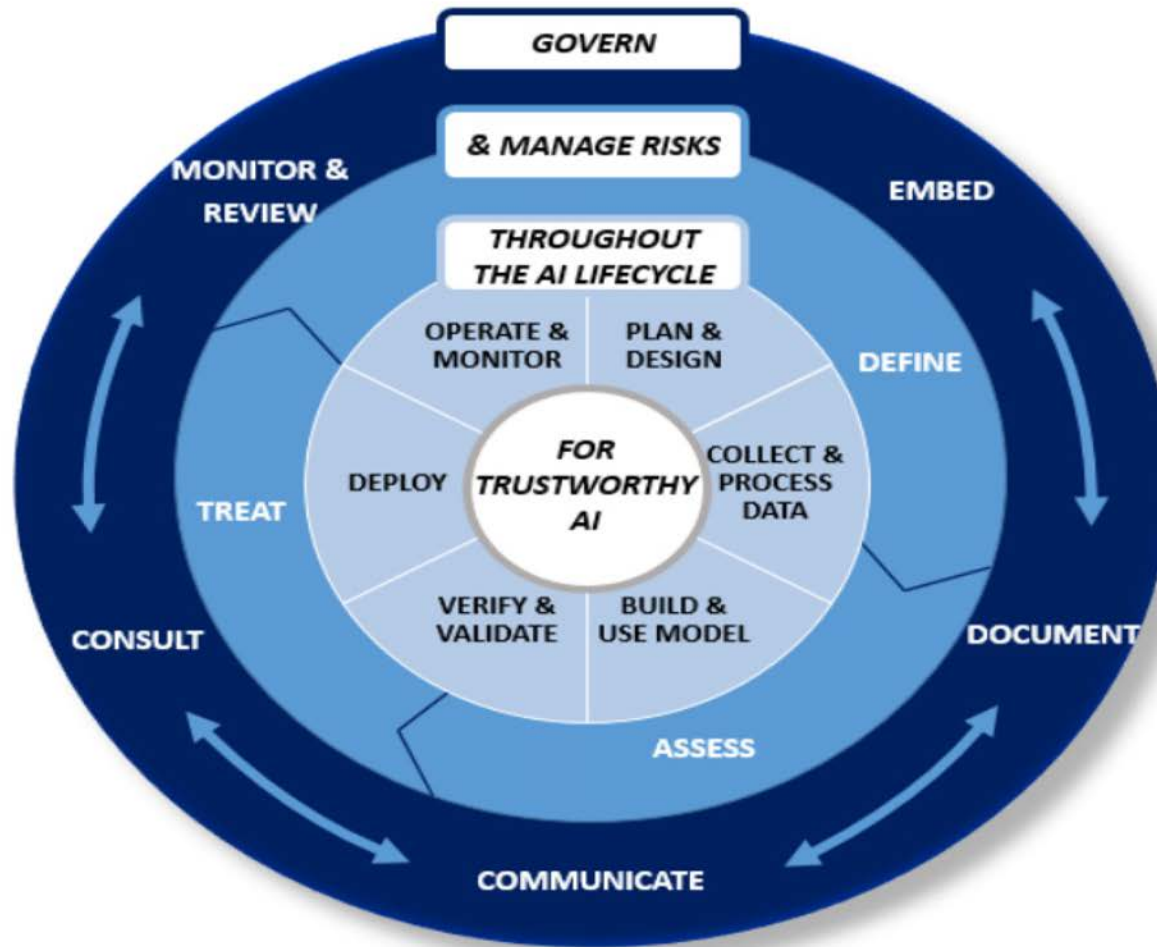
Fair with
Harmful Bias
Managed

Accountable
&
Transparent

Valid & Reliable



Governance, Risk and Compliance



OECD report



EUROPEAN
COURT
OF AUDITORS



Assessment of AI systems – Standards and frameworks

- To assess the maturity of an AI system there are security controls based on international standards and frameworks:
- ISO/IEC 42001:2023
- NIST Risk Management Framework and Playbooks
- Google secure AI Framework
- OWASP AI security overview
- OWASP AI Top risks-LLM
- OWASP AI Top 10 risks -ML



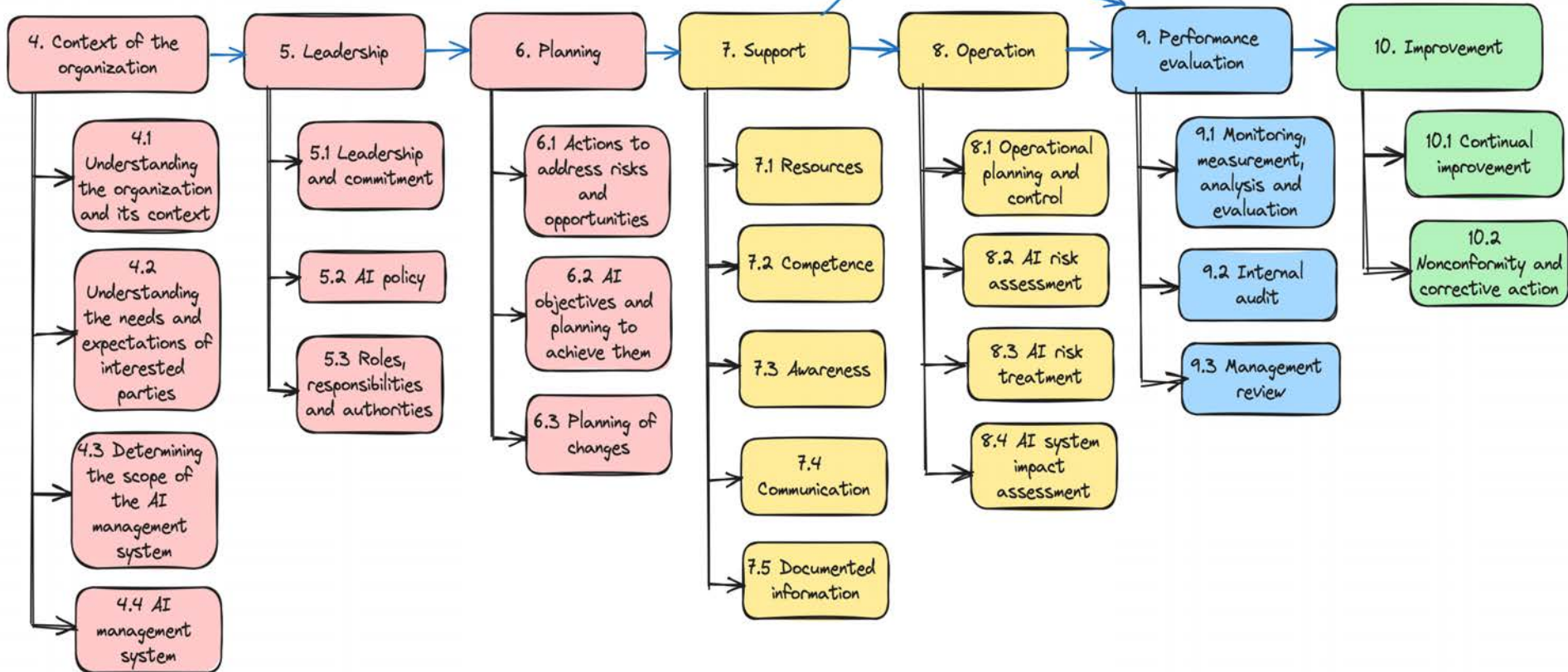
ISO 42001

Plan

Do

Check

Act



EU AI Act(1/2)

- The regulatory framework on Artificial Intelligence has the following specific objective:
 - ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values.

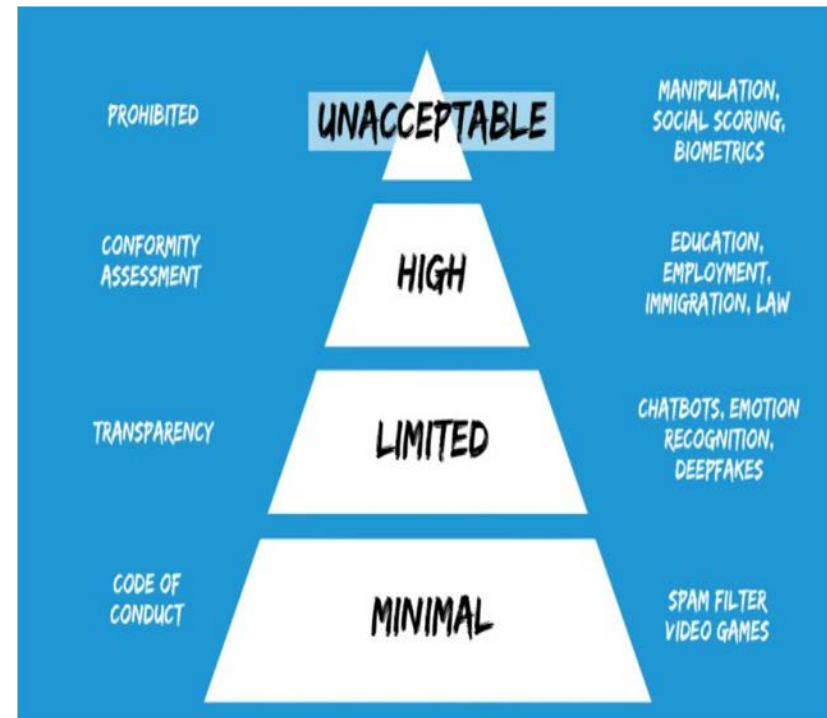
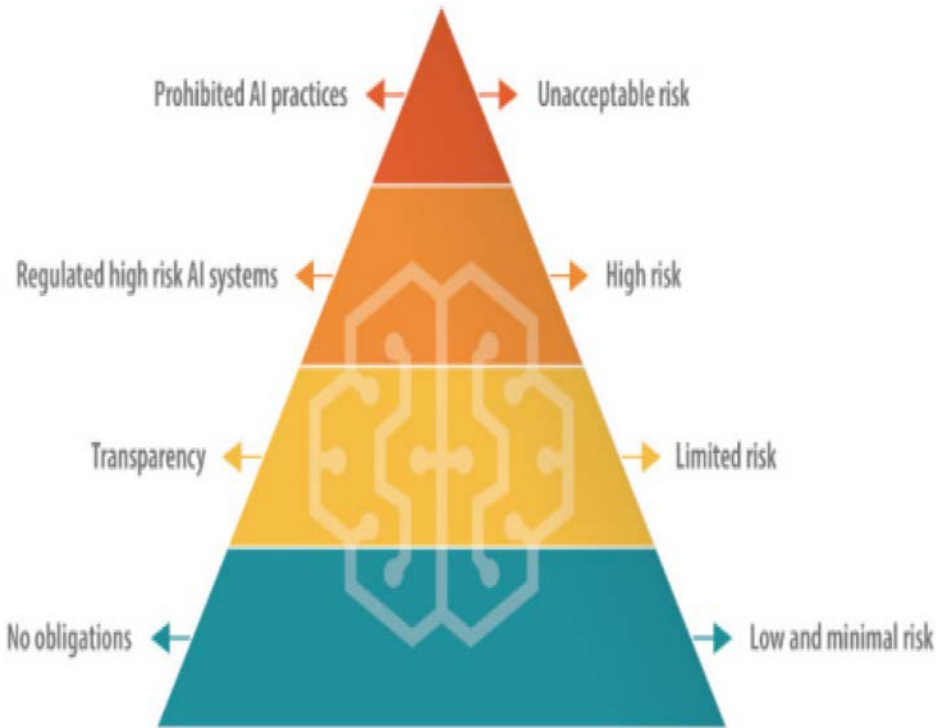


EU AI Act,
https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF



EU AI Act(2/2)

- The AI Act's primary objective is to classify AI technologies based on their potential to pose risks to society, using a risk-based approach, the greater the risk the more stringent the rules.



- On May 21st 2024 the Council gave the final green light to the first worldwide rules on AI.



Conclusions

- Users of AI need to be aware of the security and privacy risks deriving from the use of AI
- They need to be vigilant and understand the difference between real and fake pictures, videos and news.
- They need to check whether the information they receive is true, complete and accurate and not manipulated.
- Companies and organisations need to develop tools and systems with AI which
 - Follow international standards
 - Promote transparency
 - Can be audited and are not “black boxes”
 - Protect users’ privacy and personal data
 - Have been designed with security principles (security-by-design)



Thank you very much!

